



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

보건학 석사 학위논문

**Adjusting Bias of Wald Test
for Smooth Components
and Their Interactions
in Generalized Additive Model**

일반화가법모형의 평활요소 및 교호작용의
편의 보정 기법에 대한 연구

2017 년 8 월

서울대학교 보건대학원

보건학과 보건학전공

안 재 훈

Adjusting Bias of Wald Test for Smooth Components and Their Interactions in Generalized Additive Model

일반화가법모형의 평활요소 및 교호작용의
편의 보정 기법에 대한 연구

지도교수 원 성 호

이 논문을 보건학 석사 학위논문으로 제출함
2017 년 5 월

서울대학교 보건대학원

보건학과 보건학전공

안 재 훈

안재훈의 석사 학위논문을 인준함
2017 년 7 월

위 원 장	김	호 (인)
부위원장	박 태	성 (인)
위 원	원 성	호 (인)

Abstract

Generalized additive model (GAM) uses covariates based on smooth functions and can easily predict non-linear relationship between response variables and covariates. However, in spite of its flexibility, it has been known that p-values for its Wald and likelihood ratio tests do not preserve the nominal significance levels. S. N. Wood (2013) found that Wald statistics follow the mixture of weighted chi-square distribution and it has been often utilized for statistical inference. However, its performance was not carefully investigated and I found that it can lead to inflated results in certain scenarios. In my thesis, I extended his method and the proposed method was evaluated with simulation data for various hypothesis tests such as joint test for two or more smooth functions or interactions. With extensive simulations, I confirmed that the proposed method generally performs better than Wood's method. Furthermore, the proposed method was applied to the gene-by-smoking interaction association analyses. Four *SOX9*-associated SNPs were known to be associated with lung functions, and their interaction effect with *pack-years* was evaluated with Korean cohort data. Interaction test between pack-years and SNPs with linear mixed effects model was not significant but a generalized linear mixed model resulted in significant interaction, which reveals that GAM is useful for covariates with nonlinear relationships with response variables. In conclusion, GAM is useful for modeling non-linear relationship and the proposed method enables valid

statistical inferences.

Keywords: Generalized additive model (GAM), Smooth function,
Extended Wald-type test, Genome-environment wide interaction studies
(GEWIS), *SOX9*

Student Number : 2015-24078

Contents

Abstract -----	i
List of Tables -----	iv
List of Figures -----	v
 I . Introduction -----	 1
II . Method	
1. Generalized Additive Models -----	3
2. GAM as Penalized Generalized Linear Models -----	8
3. Generalized Additive Mixed Models -----	10
4. Hypothesis Test in GAM -----	11
5. Simulation Study -----	14
6. Application -----	16
III . Result	
1. Simulation -----	20
2. GEWIS -----	36
IV . Discussion -----	40
 Bibliography -----	 42
국문초록 -----	47

List of Tables

Table 1. Simulated models and their null hypotheses -----	15
Table 2. Candidate SNPs for GEWIS -----	17
Table 3. Null hypotheses of GEWIS in LMM and GAMM -----	19
Table 4. Simulated p-values which are less than 0.05 and their 95% CIs of 1 st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) -----	31
Table 5. Simulated p-values which are less than 0.05 and their 95% CIs of 2 nd null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{f}_3 = \mathbf{0}$) -----	32
Table 6. Simulated p-values which are less than 0.05 and their 95% CIs of 3 rd null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_2 = \mathbf{0}$) -----	33
Table 7. Simulated p-values which are less than 0.05 and their 95% CIs of 4 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) -----	34
Table 8. Simulated p-values which are less than 0.05 and their 95% CIs of 5 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{0}$) -----	35
Table 9. Result of effects and p-values for rs17178251 -----	37
Table 10. Result of effects and p-values for rs17765644 -----	37
Table 11. Result of effects and p-values for rs11870732 -----	38
Table 12. Result of effects and p-values for rs4793541 -----	38

List of Figures

Figure 1. QQ plots of 1 st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with TPRS -----	21
Figure 2. QQ plots of 1 st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with CRS -----	22
Figure 3. QQ plots of 2 nd null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{f}_3 = \mathbf{0}$) with TPRS	23
Figure 4. QQ plots of 2 nd null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{f}_3 = \mathbf{0}$) with CRS -	24
Figure 5. QQ plots of 3 rd null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_2 = \mathbf{0}$) with TPRS -----	25
Figure 6. QQ plots of 3 rd null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_2 = \mathbf{0}$) with CRS -----	26
Figure 7. QQ plots of 4 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with TPRS	27
Figure 8. QQ plots of 4 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with CRS -	28
Figure 9. QQ plots of 5 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{0}$) with TPRS -----	29
Figure 10. QQ plots of 5 th null hypothesis ($\mathbf{\beta}_1 = \mathbf{f}_1 = \mathbf{0}$) with CRS -----	30
Figure 11. Estimated interaction of smooth function for <i>pack-year</i> variable and rs17765644 (Table 10) -----	39

I . Introduction

Generalized additive models (GAM) (T. J. Hastie & Tibshirani, 1990) are an extension of generalized linear models to additive models which use smoothing function-based explanatory variables, and systematic component consists of linear and nonlinear parts. If I let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ be covariate matrix and $\mathbf{f} = (f_1, \dots, f_l)$ be a set of smooth functions, GAM can be defined by:

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad (1)$$

where \mathbf{y} is a response variable, $\boldsymbol{\mu} = E(\mathbf{y})$,

GAM has both parametric and nonparametric properties and can fit the nonlinear relationship between covariates and response variables. Because of these advantages, GAM is used in a variety of fields such as ecology (Fewster, Buckland, Siriwardena, Baillie, & Wilson, 2000; Suárez-Seoane, Osborne, & Alonso, 2002), epidemiology (Dominici, McDermott, Zeger, & Samet, 2002; Webster, Vieira, Weinberg, & Aschengrau, 2006), medicine (T. Hastie & Tibshirani, 1995), and public health (Hoffman et al., 2010; Vieira, Webster, Weinberg, & Aschengrau, 2009).

However, p-values for GAM tend to have an inflated type-1 error rates (S. Wood, 2006), and to overcome this, various statistical methods were suggested (Cantoni & Hastie, 2002; Crainiceanu, Ruppert, Claeskens, & Wand, 2005; Young, Weinberg, Vieira, Ozonoff, & Webster, 2011). In particular, Wood (2013)

proposed a modified Wald-type test for single smooth function that can be applied more general situations.

In my thesis, I modified Wood's method for statistical inference of multiple smoothing functions and interactions between smooth components and linear covariates, and it was compared with the existing method. Furthermore the proposed method was applied to gene-by-smoking interaction analysis with cohort data and results were compared with the linear mixed effects model.

II . Method

1. Generalized Additive Models

A generalized linear model (GLM) (McCullagh & Nelder, 1989) is an extension of classical linear model. If I let $\boldsymbol{\mu} = E(\mathbf{Y})$, \mathbf{Y} be a response variable, g be a link function, \mathbf{X} be a design matrix, GLM is given by

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. GLM assumes that the distribution of response variable \mathbf{Y} be in the exponential family. GLM shares many properties with the general linear model since the systematic component of GLM is linear combination of covariates.

A generalized additive model (GAM) is similar with GLM except that the systematic component can be non-linear. In systematic component, nonlinear part is represented by smooth functions of covariates but multiple smooth functions are assumed to be additive in the model.

Parameter estimation of smooth functions in GAM is relatively complicated as compared with GLM. If smooth functions are expressed as linear combination of unknown parameters, parameter estimation in GAM becomes equivalent to that of GLM. Let assume that q basis functions are used to express general non-linear relationship as follows:

$$f(\mathbf{x}) = \sum_{i=1}^q b_i(\mathbf{x})\beta_i,$$

where β_1, \dots, β_q are unknown parameters. Especially, the *spline* basis is widely used in GAM (Gu, 2013; Wahba, 1990), and I considered cubic spline and thin plate spline bases.

To control the flexibility of a smooth function or overfitting in GAM, penalty terms are considered for fitting the model (Parker & Rice, 1985; Wahba, 1980). One of them is the integrated square of second derivative penalty, which is given by

$$\lambda \int_0^1 [f''(x)]^2 dx \quad (2)$$

where λ is smoothing parameter. If λ is larger than 0, $f(x)$ interpolate the covariate data, and becomes overly flexible. On the other hand, if the smoothing parameter becomes 0, $f(x)$ approaches to a simple least square line and is expected to be inflexible.

Cubic regression spline

A cubic spline consists of cubic polynomial pieces and the original functions, and their first and second derivatives are continuous at the specified knots (S. Wood, 2006). This can be used in a smooth function with a single covariate. Using these properties, a cubic spline function with k knots, x_1, \dots, x_k , can be

written as

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \text{ if } x_j \leq x \leq x_{j+1}, \quad (3)$$

where $a_j^-(x) = (x_{j+1} - x)/h_j$, $a_j^+(x) = (x - x_j)/h_j$,

$$c_j^- = \left[(x_{j+1} - x)^3 / h_j - h_j(x_{j+1} - x) \right] / 6,$$

$$c_j^+ = \left[(x - x_j)^3 / h_j - h_j(x - x_j) \right] / 6 \text{ are the basis functions, } \beta_j = f(x_j),$$

$$\delta_j = f''(x_j) \text{ and } h_j = x_{j+1} - x_j.$$

In addition to the nature of the cubic spline, if the second derivative at each knots are zero, it is called a ‘*natural spline*’, and then it satisfies the following equation:

$$\mathbf{B}\boldsymbol{\delta}^- = \mathbf{D}\boldsymbol{\beta},$$

where $\boldsymbol{\delta}^- = (\delta_2, \dots, \delta_{k-1})^T$, \mathbf{B} is $(k-2) \times (k-2)$ matrix with $\mathbf{B}_{i,i} = (h_i + h_{i+1})/3$, $\mathbf{B}_{i,i+1} = \mathbf{B}_{i+1,i} = h_{i+1}/6$, and \mathbf{D} is $(k-2) \times k$ matrix with $\mathbf{D}_{i,i} = 1/h_i$, $\mathbf{D}_{i,i+1} = -1/h_i - 1/h_{i+1}$, $\mathbf{D}_{i,i+2} = 1/h_{i+1}$. Let the $k \times k$ matrix \mathbf{F} be

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}^{-1}\mathbf{D} \\ \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is a transpose of zero vector. Using the above notations, the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can be written as

$$\boldsymbol{\delta} = \mathbf{F}\boldsymbol{\beta}.$$

Hence, a cubic spline function $f(x)$ in (3) is given by

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)F_j\beta + c_j^+(x)F_{j+1}\beta \text{ if } x_j \leq x \leq x_{j+1},$$

and it can be expressed, in terms of $\boldsymbol{\beta}$, as

$$f(\mathbf{x}) = \sum_{i=1}^k b_i(\mathbf{x})\beta_i.$$

Here $b_i(\mathbf{x})$ are new basis functions in $f(\mathbf{x})$. Lancaster and Salkauskas (1986) showed that $\mathbf{D}^T \mathbf{B}^{-1} \mathbf{D}$ is the penalty matrix for cubic spline basis, in other words,

$$\int f''(\mathbf{x})^2 d\mathbf{x} = \boldsymbol{\beta}^T \mathbf{D}^T \mathbf{B}^{-1} \mathbf{D} \boldsymbol{\beta}.$$

Thin plate regression spline

A thin plate spline (Duchon, 1977) is the other solution of smooth function in GAM. The thin plate splines have several advantages as compared with the cubic splines (S. N. Wood, 2003). Unlike cubic splines, the smooth function using thin plate splines can be expressed with multiple covariates. In addition, there is no need to choose knots in thin plate spline.

In order to estimate the thin plate spline function, the following equations need to be minimized:

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda J_{md}(\mathbf{f}), \quad (4)$$

where $\mathbf{f} = (f(x_1), f(x_2), \dots, f(x_n))^T$, λ is smoothing parameter and $J_{md}(\mathbf{f})$ is defined as

$$J_{md} = \int \cdots \int_{\mathbb{R}^d} \sum_{v_1 + \cdots + v_d = m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial x_1^{v_1} \cdots \partial x_d^{v_d}} \right)^2 dx_1 \cdots dx_d.$$

With the technical restriction $2m > d$, thin plate spline function in (4) can be written as

$$f(\mathbf{x}) = \sum_{i=1}^n \delta_i \eta_{md}(\|\mathbf{x} - x_i\|) + \sum_{j=1}^M \alpha_j \phi_j(\mathbf{x}), \quad (5)$$

where $M = \binom{m+d-1}{d}$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ and $\boldsymbol{\alpha} = (\alpha_j, \dots, \alpha_M)$ are vectors of unknown parameters, especially $\boldsymbol{\delta}$ is subject to the linear constraints $\mathbf{T}^T \boldsymbol{\delta} = \mathbf{0}$ where $T_{ij} = \phi_j(x_i)$. In addition, function ϕ_j of M in (5) spans the space of functions which J_{md} becomes null space. Function η_{md} in (5) is defined as

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1} \pi^{d/2} (m-1)! (m-d/2)!} r^{2m-d} \log(r) & d \text{ even} \\ \frac{\Gamma(d/2 - m)}{2^{2m} \pi^{d/2} (m-1)!} r^{2m-d} & d \text{ odd.} \end{cases}$$

For reducing computation cost, (4) is rewritten as

$$\|\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}^T \mathbf{E} \boldsymbol{\delta} \text{ subject to } \mathbf{T}^T \boldsymbol{\delta} = \mathbf{0}, \quad (6)$$

where matrix \mathbf{E} has elements $E_{ij} \equiv \eta_{md}(\|x_i - x_j\|)$. Let $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ by Eigen-decomposition, then the first k columns of \mathbf{U} is denoted by \mathbf{U}_k , \mathbf{D}_k indicates $k \times k$ submatrix in \mathbf{D} and $\delta_k = \mathbf{U}_k^T \boldsymbol{\delta}$. Using above notations, (6) can be represented as

$$\|\mathbf{y} - \mathbf{U}_k \mathbf{D}_k \delta_k - \mathbf{T} \boldsymbol{\alpha}\|^2 + \lambda \delta_k^T \mathbf{D}_k \delta_k \text{ subject to } \mathbf{T}^T \mathbf{U}_k \delta_k = 0,$$

and find \mathbf{U}_k and \mathbf{D}_k by Lanczos algorithm (Demmel, 1997).

2. GAM as Penalized Generalized Linear Models

Recall that GAM consists of a linear predictor terms and some smooth functions.

To estimate GAM, it is convenient to express the smooth function in a form similar to a linear predictor with appropriate bases. In this framework, a smooth function can be represented as

$$f_i(x_k) = \sum_{j=1}^{q_i} \beta_{ij} b_{ij}(x_k)$$

where $\mathbf{b}_i = (b_{i1}, \dots, b_{iq_i})^T$ is a vector of basis functions which is chosen for smooth function f_i and $\tilde{\boldsymbol{\beta}}_i = (\beta_{i1}, \dots, \beta_{iq_i})^T$ is unknown parameter for f_i which need to be estimated in GAM (S. Wood, 2006). If an element of matrix $\tilde{\mathbf{X}}_i$ is expressed by given basis functions, then

$$\mathbf{f}_i = (f_i(x_{k1}), \dots, f_i(x_{kn})) = \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\beta}}_i, \quad (7)$$

and $\tilde{\mathbf{X}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ are re-parameterized for centering constraint by column orthogonal matrix \mathbf{Z} which satisfies $\mathbf{1}^T \tilde{\mathbf{X}}_i \mathbf{Z} = 0$. Using \mathbf{Z} in (7),

$$\mathbf{f}_i = \tilde{\mathbf{X}}_i \mathbf{Z}^T \mathbf{Z} \tilde{\boldsymbol{\beta}}_i = \mathbf{X}_i \boldsymbol{\beta}_i \quad (8)$$

where $\mathbf{X}_i = \tilde{\mathbf{X}}_i \mathbf{Z}^T$, $\boldsymbol{\beta}_i = \mathbf{Z} \tilde{\boldsymbol{\beta}}_i$ satisfies the centering constraint.

Hence, with above notation, GAM can be expressed as GLM:

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta} \quad (9)$$

where $\mathbf{X} = (\mathbf{X}^*, \mathbf{X}_1, \mathbf{X}_2, \dots)$, $\boldsymbol{\beta}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots)$ and $\mathbf{X}^* \boldsymbol{\theta}$ is linear predictor term in (1). In addition, penalty term (2) in model also can be represented by above notation as

$$\sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta} \quad (10)$$

where $\mathbf{S}_j = \mathbf{Z}^T \tilde{\mathbf{S}}_j \mathbf{Z}$ and $\tilde{\mathbf{S}}_j$ is a matrix of known coefficients for penalty as $\tilde{\boldsymbol{\beta}}^T \tilde{\mathbf{S}}_j \tilde{\boldsymbol{\beta}}$. With this approach, the unknown parameters $\boldsymbol{\beta}$ can be estimated by maximizing penalized likelihood l_p :

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta},$$

where $l(\boldsymbol{\beta})$ is likelihood of model and λ_j is already estimated by generalized cross validation or maximum likelihood for smoothing parameter.

Degrees of freedom for GAM also need to consider. In our approach for GAM by penalized GLM, applying the *effective degrees of freedom* in penalized model (Zou, Hastie, & Tibshirani, 2007) makes sense which is defined as

$$\text{edf} \equiv \text{tr} \left(\mathbf{X} (\mathbf{X}^T \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^T \right), \quad (11)$$

where \mathbf{X} in (9) and $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$ in (10).

3. Generalized Additive Mixed Models

The approach of GAM to penalized GLM as above is the same in Generalized additive mixed models (GAMM) (Lin & Zhang, 1999; S. Wood, 2006). The structure of GAMM is written as

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots + \mathbf{Z}_i \mathbf{b}, \quad (12)$$

where \mathbf{Z}_i is the model matrix of variables to which the random effect is applied, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$ and the rest part of structure is same as (1). Using the approach in (9), (12) can be expressed as generalized linear mixed models (GLMM) (Stroup, 2012):

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}.$$

In this framework, parameters in GAMM can be estimated as in GAM and GLMM: the bases of smooth function and smoothing parameter for penalty in GAM, the covariance matrix of random effects in GLMM. Although there are various methods for estimation in linear mixed model, maximum likelihood (ML) and restricted maximum likelihood (REML) (Patterson & Thompson, 1971) are most commonly used (Breslow & Clayton, 1993; Pinheiro & Bates, 2000; Searle, Casella, & McCulloch, 1992).

4. Hypothesis Test in GAM

For hypothesis testing, distribution of estimated parameters has to be considered. Using weighted least square method with penalties, $\hat{\beta}$ is given by

$$\hat{\beta} = (X^T W X + S)^{-1} X^T W y$$

where $W^{-1}\phi$ is covariance matrix of response variable y , and its (frequentist) covariance matrix is

$$V_{\hat{\beta}} = (X^T W X + S)^{-1} X^T W X (X^T W X + S)^{-1} \phi.$$

Although $\hat{\beta}$ is a biased estimator because of penalty terms, $\hat{\beta}_j$ which is subset of $\hat{\beta}$ approximately follows normal distribution under the null hypothesis $\beta_j = 0$ (S. Wood, 2006) :

$$\hat{\beta}_j \sim N(\mathbf{0}, V_{\hat{\beta}_j}),$$

and using $r = \text{rank}(V_{\hat{\beta}_j})$, Wald test statistic and its distribution can be represented as

$$\hat{\beta}_j^T V_{\hat{\beta}_j}^{r-} \hat{\beta}_j \sim \chi_r^2 \quad (13)$$

where $V_{\hat{\beta}_j}^{r-}$ is rank r pseudoinverse matrix.

The problem of approximation in (13) is that p-values tend to be liberal (S. Wood, 2006), and so our test results become unreliable.

There is a similar problem in likelihood ratio test (LRT) for GAM. Considering the definition of effective degree of freedom in (11), it is natural

for the likelihood ratio statistic to follow the approximated distribution as

$$\text{LR} = 2 \left(l(\hat{\beta}_1) - l(\hat{\beta}_0) \right) \sim \chi^2_{\text{edf1}-\text{edf0}} \quad (14)$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are subsets of $\hat{\beta}$ ($\hat{\beta}_0$ is nested in $\hat{\beta}_1$), edf1 and edf0 are effective degrees of freedom of models for estimation $\hat{\beta}_1$ and $\hat{\beta}_0$, respectively. However, LRT with (14) is also inaccurate, especially because of penalty terms in models.

S. N. Wood (2013) proposed the modified Wald-type test for null hypothesis $f_i = 0$ in GAM with thin plate regression spline using Eigen-decomposition of $\mathbf{V}_{f_i}^{r-}$ with alternative definition of effective degree of freedom r as

$$r = \text{tr}(2\mathbf{A} - \mathbf{A}^2)$$

where \mathbf{A} is diagonal matrix of Eigen-values λ_j of

$$\mathbf{V}_{f_i} = \mathbf{X}_i \mathbf{V}_{\beta_i} \mathbf{X}_i^T \quad (15)$$

in (8) and (13). Briefly, if r is not an integer,

$$\mathbf{V}_{f_i}^{r-} = \mathbf{U} \begin{pmatrix} \lambda_1^{-1} & & & \\ & \ddots & & \\ & & \lambda_{k-2}^{-1} & \\ & & & \mathbf{B} \\ & & & & \mathbf{0} \end{pmatrix} \mathbf{U}^T$$

for \mathbf{U} is the matrix with each column is Eigen-vector of \mathbf{V}_{f_i} , $k = [r] + 1$ and

$$\mathbf{B} = \tilde{\mathbf{A}} \tilde{\mathbf{B}} \tilde{\mathbf{A}} \quad \text{where} \quad \tilde{\mathbf{A}} = \begin{pmatrix} \lambda_{k-1}^{-1/2} & 0 \\ 0 & \lambda_k^{-1/2} \end{pmatrix}, \quad \tilde{\mathbf{B}} = \begin{pmatrix} 1 & \rho \\ \rho & \nu \end{pmatrix}, \quad \nu = r - k + 1 \quad \text{and}$$

$\rho = (\nu(1 - \nu)/2)^{1/2}$. The modified Wald statistic T_r is defined by

$$T_r = \boldsymbol{\delta}_1^T \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2^T \tilde{\mathbf{B}} \boldsymbol{\delta}_2$$

where $\boldsymbol{\delta}_1 = (d_1, \dots, d_{k-2})^T$, $\boldsymbol{\delta}_2 = (d_{k-1}, d_k)^T$ and $\mathbf{d} = (d_1, \dots, d_{q_i})^T = \mathbf{A}^{-1/2} \mathbf{A} \mathbf{U}^T \mathbf{y}$. Therefore, T_r follows mixture chi-square distribution as

$$T_r \sim \chi_{k-2}^2 + v_1 \chi_1^2 + v_2 \chi_1^2$$

where v_1 and v_2 are Eigen-values of $\tilde{\mathbf{B}}$ and calculate p-values using approximation by Liu, Tang, and Zhang (2009).

In this thesis, I evaluated the empirical type-1 error rates of Wald-type test by S. N. Wood (2013) when (i) $f_i = f_j = 0$; (ii) $f_i = f_j = f_k = 0$; (iii) $\theta_i = f_j = 0$ where θ_i is parameter in linear predictor term in (1); $\theta_j = f_k = f_{k,j} = 0$ where $f_{k,j}$ is an interaction smooth function with x_k and x_j which is column of \mathbf{X}^* in (9); and (v) $f_{k,j_1} = f_{k,j_2} = 0$ where f_{k,j_i} is an interaction smooth function. Recall that the covariance matrix of f_i is given by (15), and consider the covariance matrix for null hypothesis (1) as

$$\mathbf{V}_{f_i f_j} = \mathbf{X}_+ \mathbf{V}_{\boldsymbol{\beta}_+} \mathbf{X}_+^T$$

where $\mathbf{X}_+ = (\mathbf{X}_i, \mathbf{X}_j)$ and $\mathbf{V}_{\boldsymbol{\beta}_+}$ is covariance matrix of $\boldsymbol{\beta}_+ = (\beta_i^T, \beta_j^T)^T$. $\mathbf{V}_{\boldsymbol{\beta}_+}$ takes into account correlations between all parameters from each smooth function and from the different smooth functions, as \mathbf{V}_{β_i} in (8) includes information of correlations between parameters from single smooth function. Other covariance matrix for null hypotheses (2) to (5) can also be expressed in a similar way.

In addition, some modifications apply to Wald-type test in S. N. Wood (2013). Firstly, re-define ρ in $\tilde{\mathbf{B}}$ as $\rho' = \sqrt{v(1-v)}$, and secondly, use scaled F distribution approximation to mixture chi-square distribution for calculating p-values (Wu & Lin, 2016). These are also applied to cubic spline regression.

5. Simulation Study

The extension of Wald-type test can be evaluated through simulated data. For each of hypotheses, these 5 methods are compared each other: (i) Wald-type test method of S. N. Wood (2013); (ii) using ρ' instead of ρ ; (iii) using scaled F distribution approximation; applying modifications both (ii) and (iii); (v) approximated Wald test in (13).

Number of subjects were assumed to have n subjects, and 8 covariates were generated. First 5 covariates with linear effect on responses are denoted by x_0 to x_4 , and are generated as follows: $x_0 \sim N(5, 2)$; $x_1 \sim \text{Poisson}(15)$; $x_2 \sim N(3, 1)$; $x_3 = \sin(t)$, $t \sim N(0, 6)$; x_4 is an integer randomly selected from -100 to 100. y is generated by $\sum_{j=0}^4 x_j + \varepsilon$, $\varepsilon \sim N(2, 3)$. Last 3 covariates are denoted by s_1 to s_3 , and were used as components in smooth function terms and their interactions. They were simulated from $N(3, 1)$, $U(0, 5)$, $N(-1, 2)$, respectively.

Table 1 shows five models to verify the performance of the extended tests and their null hypotheses. For each models, 10,000 replicate data which have 100, 500 and 5,000 observations were simulated. P-values of each null hypotheses are compared with quantiles from $U(0,1)$ using log-scaled quantile-quantile (QQ) plot. In addition, since statistical analysis of many studies uses a 0.05 significance level, I evaluated the simulated p-values which are less than 0.05 and their 95% confidence intervals using bootstrap method with re-sampling of 10,000 (DiCiccio & Efron, 1996; Efron & Tibshirani, 1986). All simulation studies were conducted with mgcv 1.8-17 package (Team, 2014; S. N. Wood, 2001) of R (version 3.3.2).

Table 1. Simulated models and their null hypotheses

No.	Models	Null hypotheses
1	$y = X^*\theta + f_1(s_1) + f_2(s_2)$	$f_1 = f_2 = 0$
2	$y = X^*\theta + f_1(s_1) + f_2(s_2) + f_3(s_3)$	$f_1 = f_2 = f_3 = 0$
3	$y = X^*\theta + \beta_1 s_1 + f_2(s_2)$	$\beta_1 = f_2 = 0$
4	$y = X^*\theta + \beta_1 s_1 + f_1(s_2) + f_2(s_2)s_1$	$\beta_1 = f_1 = f_2 = 0$
5	$y = X^*\theta + \beta_1 s_1 + f_1(s_2)s_1 + f_1(s_3)s_1$	$\beta_1 = f_1 = 0$

6. Application

Data description

From Korea Associated Resource (KARE) project, data for Korean Genome Epidemiology Study (KoGES) were collected of which Ansung and Ansan cohorts. Variants and individuals, which meet the following conditions, were excluded from association analyses: the missing genotype call rates of variants are larger than 0.05; minor allele frequencies (MAFs) are less than 0.05; Hardy-Weinberg equilibrium (HWE) p values are less than 10^{-5} ; participants with missing genotype call rates larger than 0.05 or with gender inconsistencies. After quality control procedures, 8,773 participants who aged 40 to 69 and 310,515 variants are remained in data.

8,534 participants comprising 4,001 men and 4,533 women were repeatedly taken spirometry test at maximum 3 times for every 2 years, and so, a total of 19,557 observations are used for analyses. FEV₁ (Volume that has been exhaled at the end of the first second of forced expiration, see National (2010)) which measured from each test is considered a response variable in the model. Environmental variables are both pack-years and smoking status that are collected from questionnaire. Participants are divided into 2 groups: 4,926 never-smokers and 3,608 ex- or current-smokers. According to Rockich et al. (2013) and Li et al. (2015), SNPs associated with *SOX9* gene are considered for GEWIS (Genome-Environment Wide Interaction Studies): rs17178251,

rs17765644, rs11870732, rs4793541. Details of these SNPs are represented in Table 2.

Table 2. Candidate SNPs for GEWIS

SNP	Chr	Alleles	MAFs	HWE p-values
rs17178251	17	C/T	0.384	0.604
rs17765644	17	G/C	0.383	0.572
rs11870732	17	G/A	0.384	0.636
rs4793541	17	C/T	0.391	0.324

Statistical modeling

Two models are fitted for GEWIS: LMM and GAMM. The structure of both models is consistent except for a few parts. Firstly, LMM is heteroskedastic model for *smoking-status* variable but GAMM is weighted model of which weights are inverse of variances for same model with *smoking-status* grouped data. Homoskedastic GAMM is also considered. Secondly, the smooth function with 100 bases for *pack-year* variable is considered in GAMM which is not in LMM. Each models are as follows:

$$\begin{aligned}
 \text{[LMM]} \quad y_{gij} = & \beta_0 + \beta_1 age_{gi} + \beta_2 sex_{gi} + \beta_3 BMI_{ij} + \beta_4 height_{gij} + \\
 & \beta_5 time_{gij} + \beta_6 sex_i \cdot age_i + \beta_7 pack - year_{ij} + \beta_8 smoking - status_i + \\
 & \beta_9 age_i \cdot smoking - status_i + \beta_{10} sex_i \cdot smoking - status_i + \\
 & \beta_{11} height_{ij} \cdot smoking - status_i + \beta_{12} time_{ij} \cdot smoking - status_i +
 \end{aligned}$$

$$\begin{aligned} & \beta_{13}SNP_i + \beta_{14}SNP_i \cdot smoking - status_i + \beta_{15}SNP_i \cdot pack - year_{ij} + \\ & \sum_{k=1}^{10} \tau_k PC_i^k + \tau_{11} PC_i^1 \cdot smoking - status_i + b_{gi} + \varepsilon_{gij}, \\ & \varepsilon_{gij} \sim MVN(0, \Sigma_g), \quad b_{gi} \sim iid \quad N(0, \sigma_g^2), \end{aligned}$$

$$\begin{aligned} \text{[GAMM]} \quad y_{gij} = & \beta_0 + \beta_1 age_{gi} + \beta_2 sex_{gi} + \beta_3 BMI_{ij} + \beta_4 height_{gij} + \\ & \beta_5 time_{gij} + \beta_6 sex_i \cdot age_i + f_7(pack - year_{ij}) + \beta_8 smoking - status_i + \\ & \beta_9 age_i \cdot smoking - status_i + \beta_{10} sex_i \cdot smoking - status_i + \\ & \beta_{11} height_{ij} \cdot smoking - status_i + \beta_{12} time_{ij} \cdot smoking - status_i + \\ & \beta_{13} SNP_i + \beta_{14} SNP_i \cdot smoking - status_i + f_{15}(pack - year_{ij}) \cdot SNP_i + \\ & \sum_{k=1}^{10} \tau_k PC_i^k + \tau_{11} PC_i^1 \cdot smoking - status_i + b_{gi} + \varepsilon_{gij}, \\ & \varepsilon_{gij} \sim MVN(0, \Sigma_g), \quad b_{gi} \sim iid \quad N(0, \sigma_g^2), \end{aligned}$$

where PC^k is first 10 principle component score, i, j, g indicate participants, repeated measurement and *smoking-status* group, respectively.

For both models, p-values of each null hypotheses about SNP and its interaction with environmental variables are compared. [Table 3] shows the null hypotheses in each models. LMM is fitted by SAS and tested using F-statistic adjusted by Kenward-Roger approximation (Kenward & Roger, 1997).

Table 3. Null hypotheses of GEWIS in LMM and GAMM

Null hypotheses	LMM	GAMM
Total effects of SNP	$\beta_{13} = \beta_{14} = \beta_{15} = 0$	$\beta_{13} = \beta_{14} = f_{15} = 0$
Effects of SNP	$\beta_{13} = 0$	$\beta_{13} = 0$
Interaction effects of SNP× <i>smoking-status</i>	$\beta_{14} = 0$	$\beta_{14} = 0$
Interaction effects of SNP× <i>pack-year</i>	$\beta_{15} = 0$	$f_{15} = 0$

III. Result

1. Simulation

The simulation results of QQ plot are represented in Figure 1 to Figure 10 below. Each figure has 15 QQ plots and in each column, they are for simulated data with 100, 500, and 5000 observations, respectively. Plots for the five test methods in Table 1 are arranged in row order for all figures. The results of simulated p-values which are less than 0.05 and their 95% CIs of all simulated data are represented in Table 4 to Table 9.

In all simulated models, the proposed Wald-type test method and its modifications show significantly better performance than approximated Wald test (13) in almost simulated data at the 0.05 nominal significance level although modified test methods are not significantly different from that of the originally proposed by S. N. Wood (2013). For simulated data with $n = 100$ observations, proposed tests with second model in Table 1 are not perfect either (Figure 3 and Table 5).

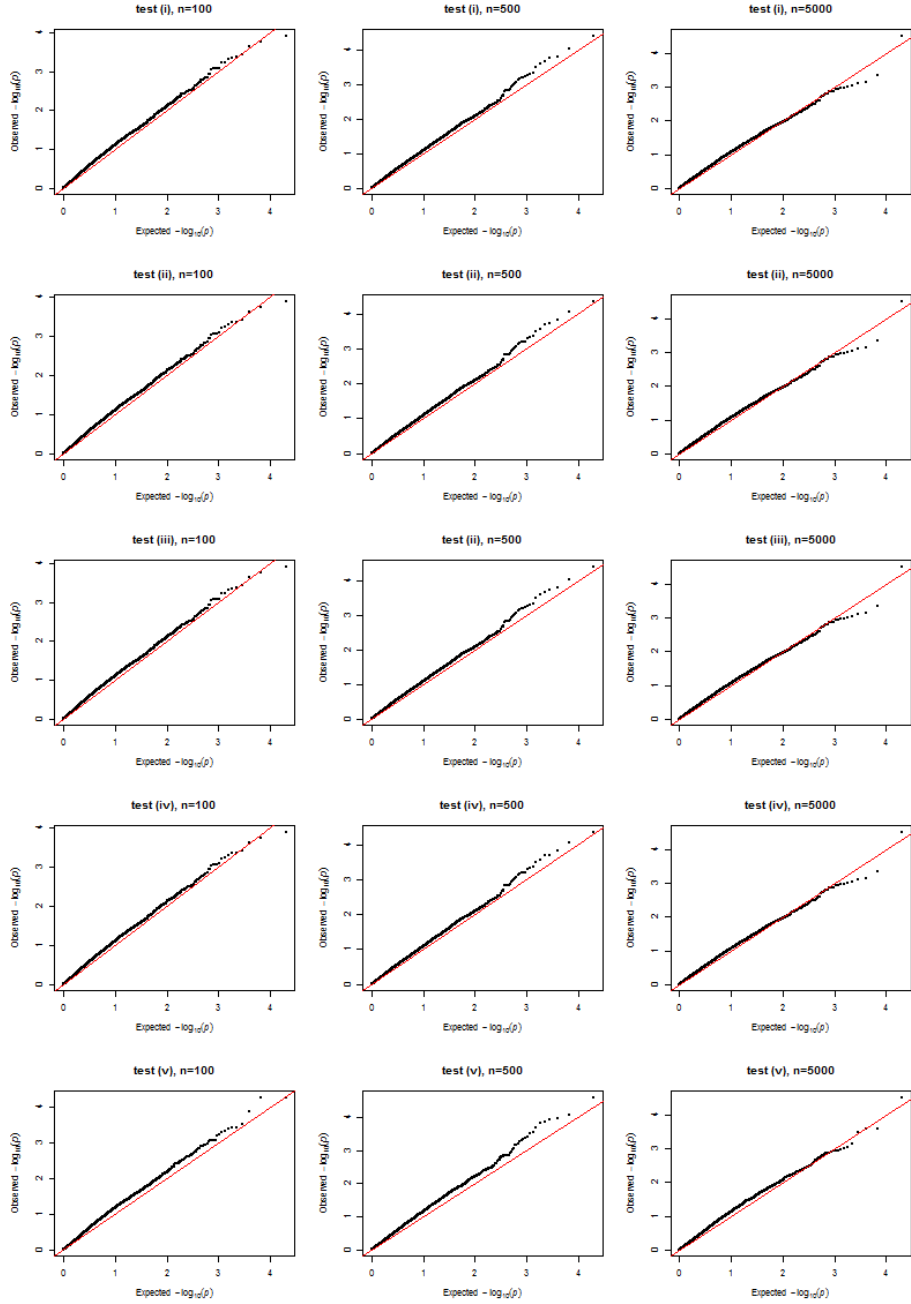


Figure 1. QQ plots of 1st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with TPRS

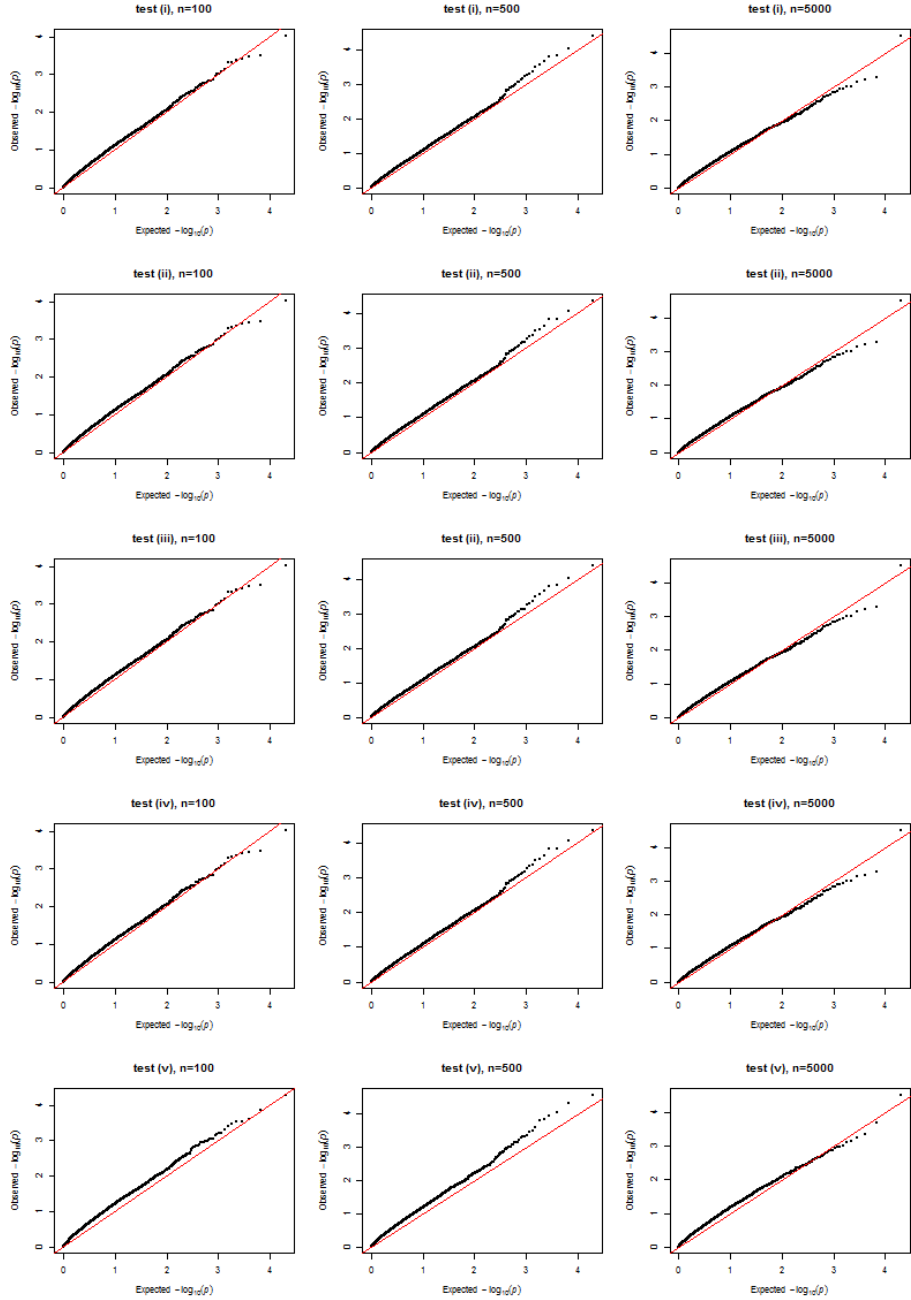


Figure 2. QQ plots of 1st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$) with CRS

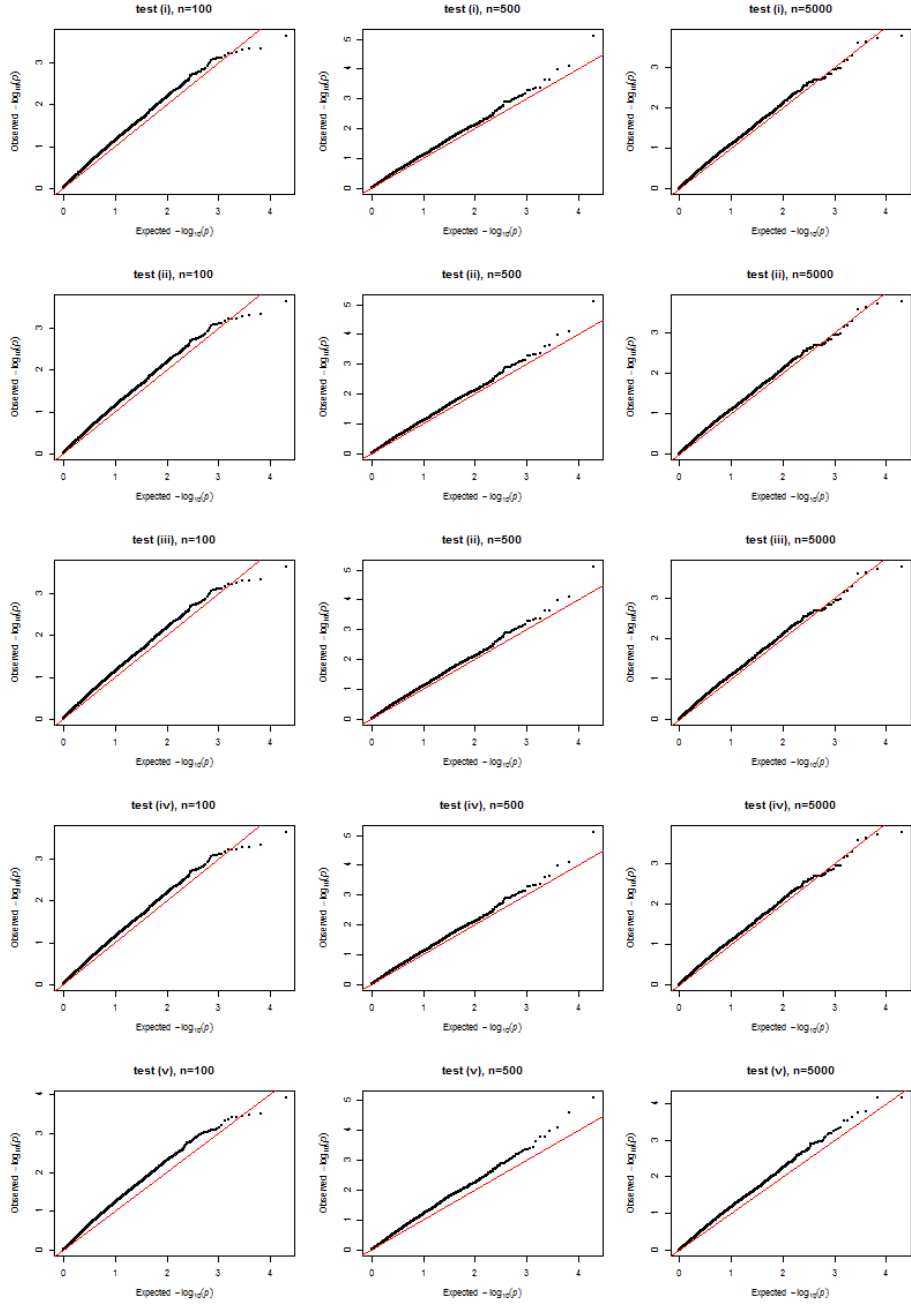


Figure 3. QQ plots of 2^{nd} null hypothesis ($f_1 = f_2 = f_3 = 0$) with TPRS

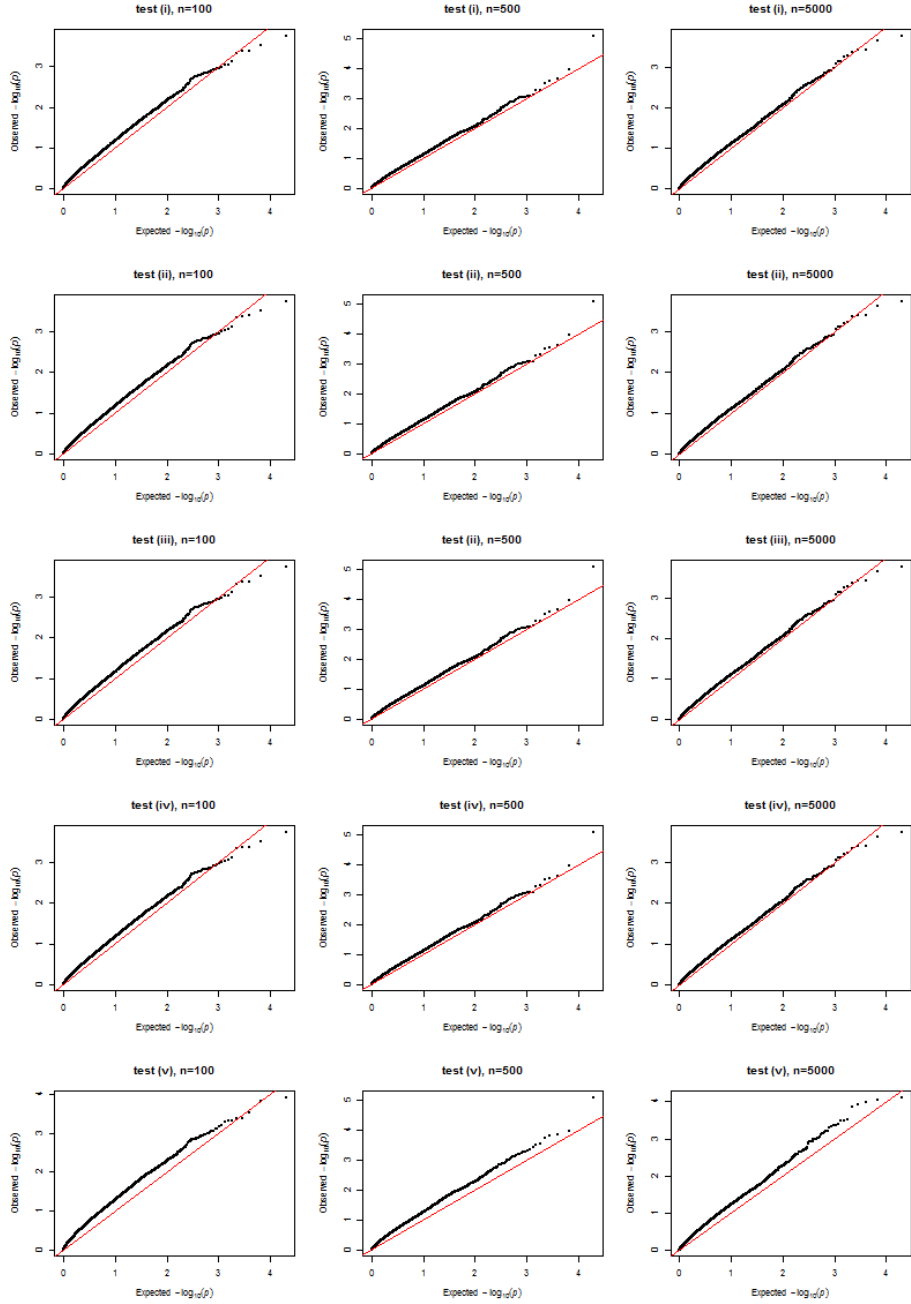


Figure 4. QQ plots of 2nd null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{f}_3 = \mathbf{0}$) with CRS

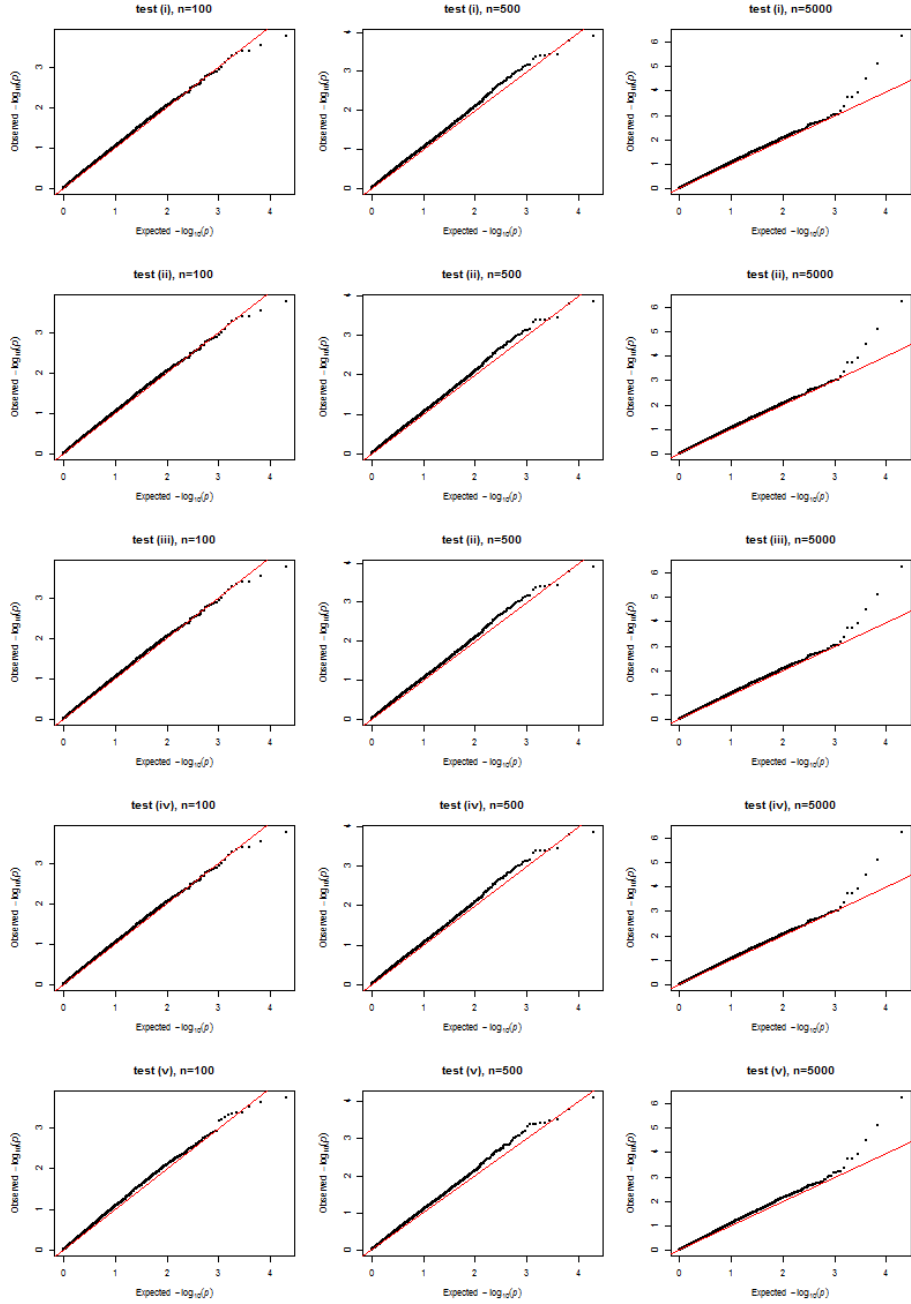


Figure 5. QQ plots of 3rd null hypothesis ($\beta_1 = f_2 = 0$) with TPRS

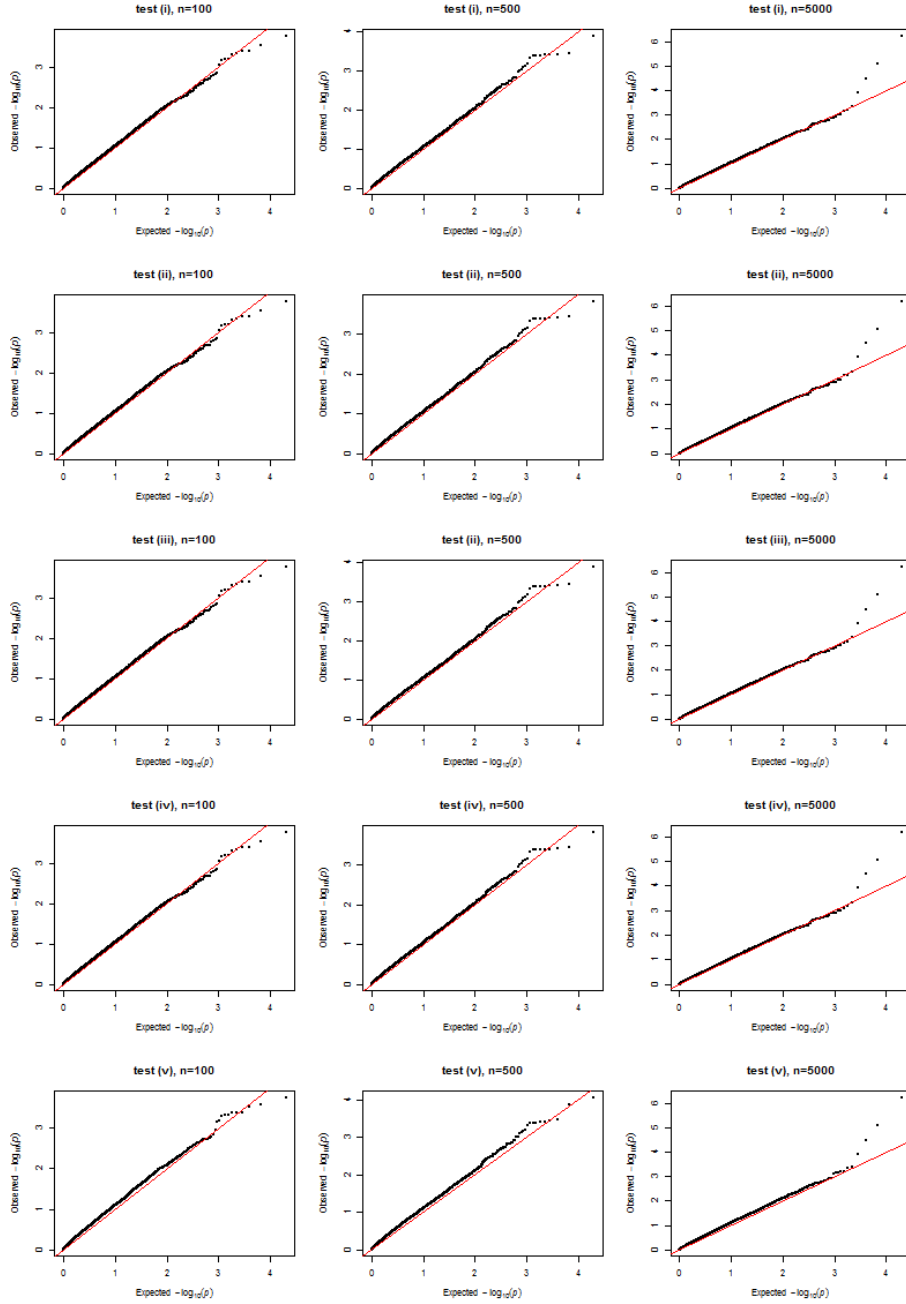


Figure 6. QQ plots of 3rd null hypothesis ($\beta_1 = f_2 = 0$) with CRS

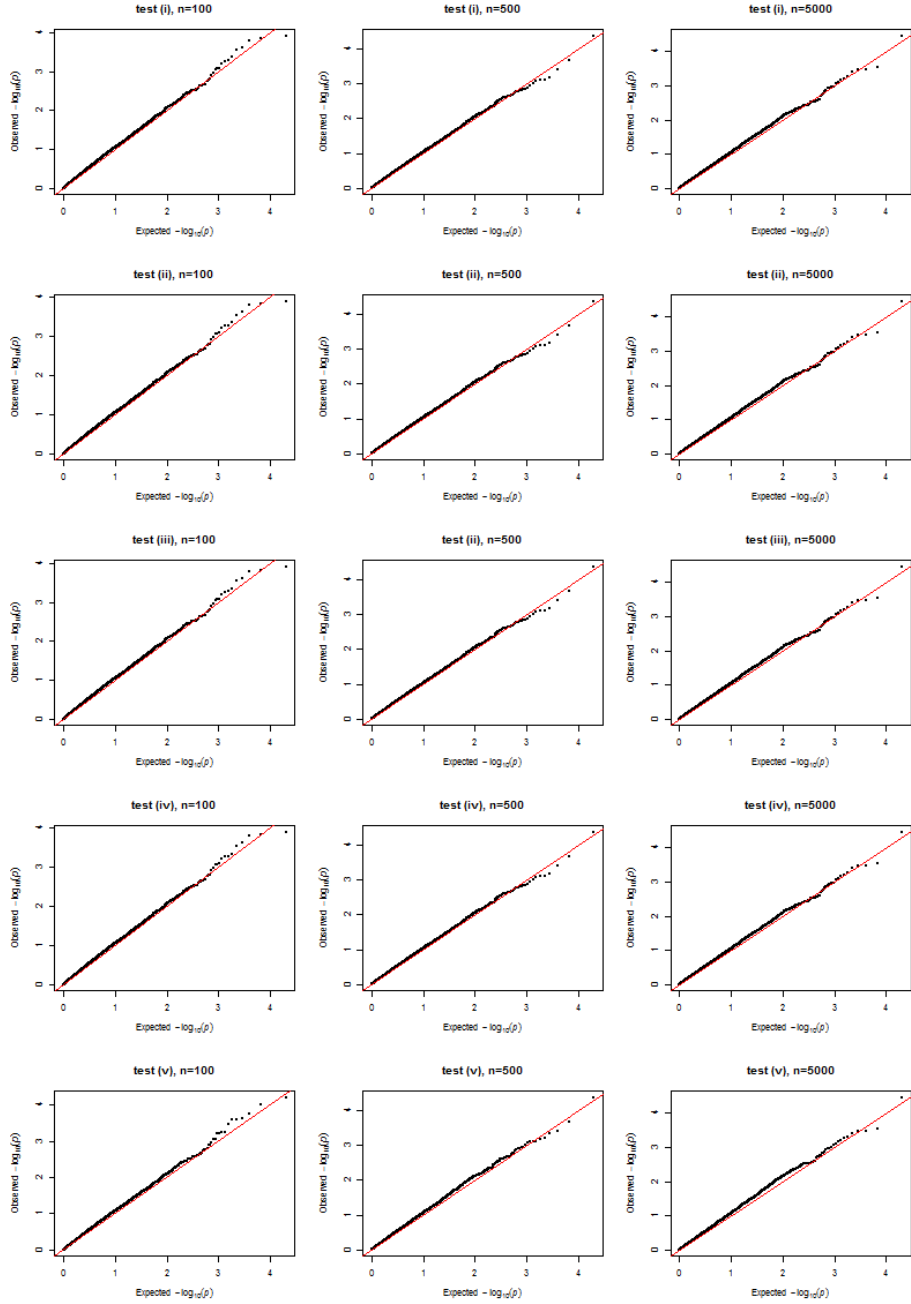


Figure 7. QQ plots of 4th null hypothesis ($\beta_1 = f_1 = f_2 = 0$) with TPRS

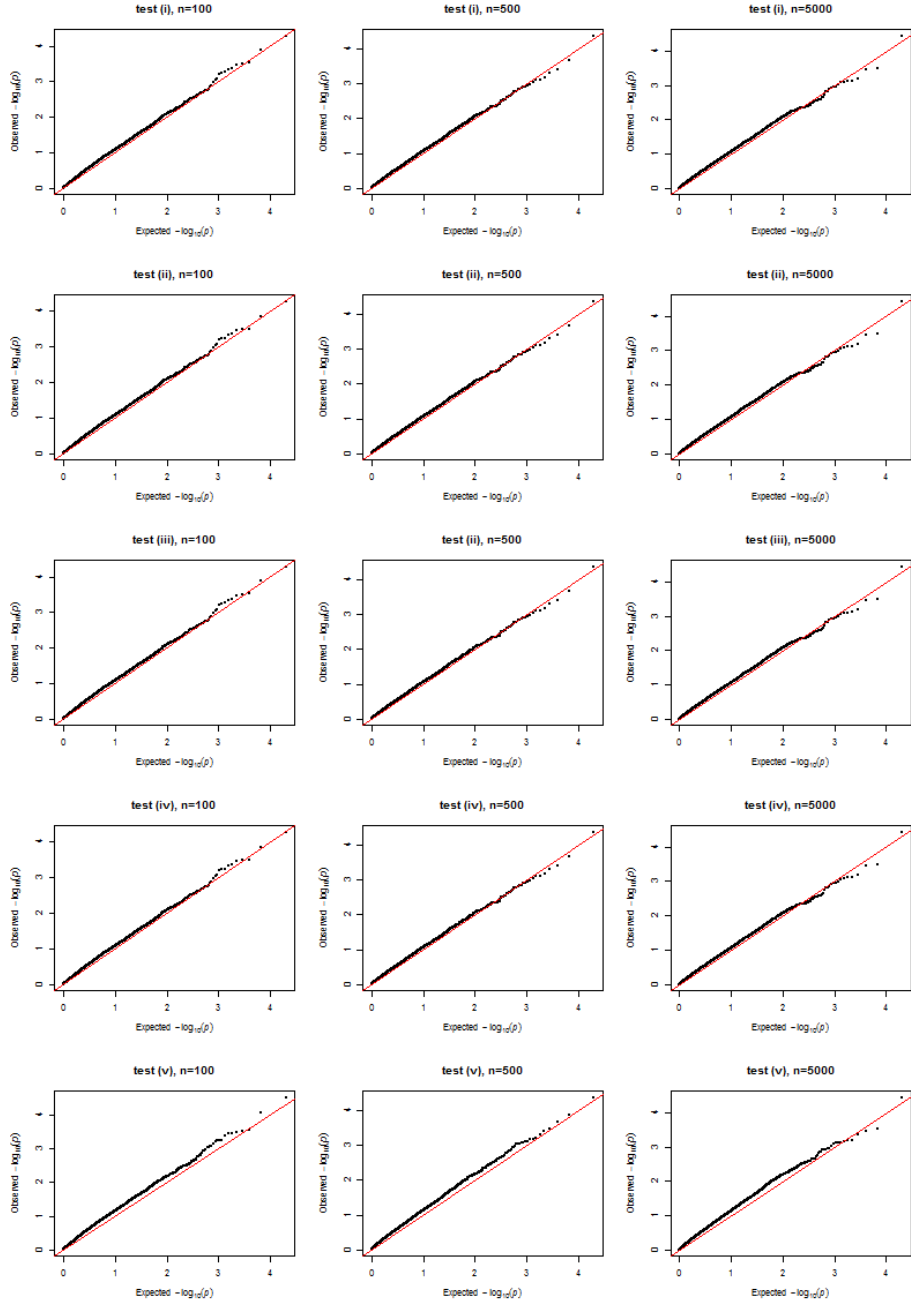


Figure 8. QQ plots of 4th null hypothesis ($\beta_1 = f_1 = f_2 = 0$) with CRS

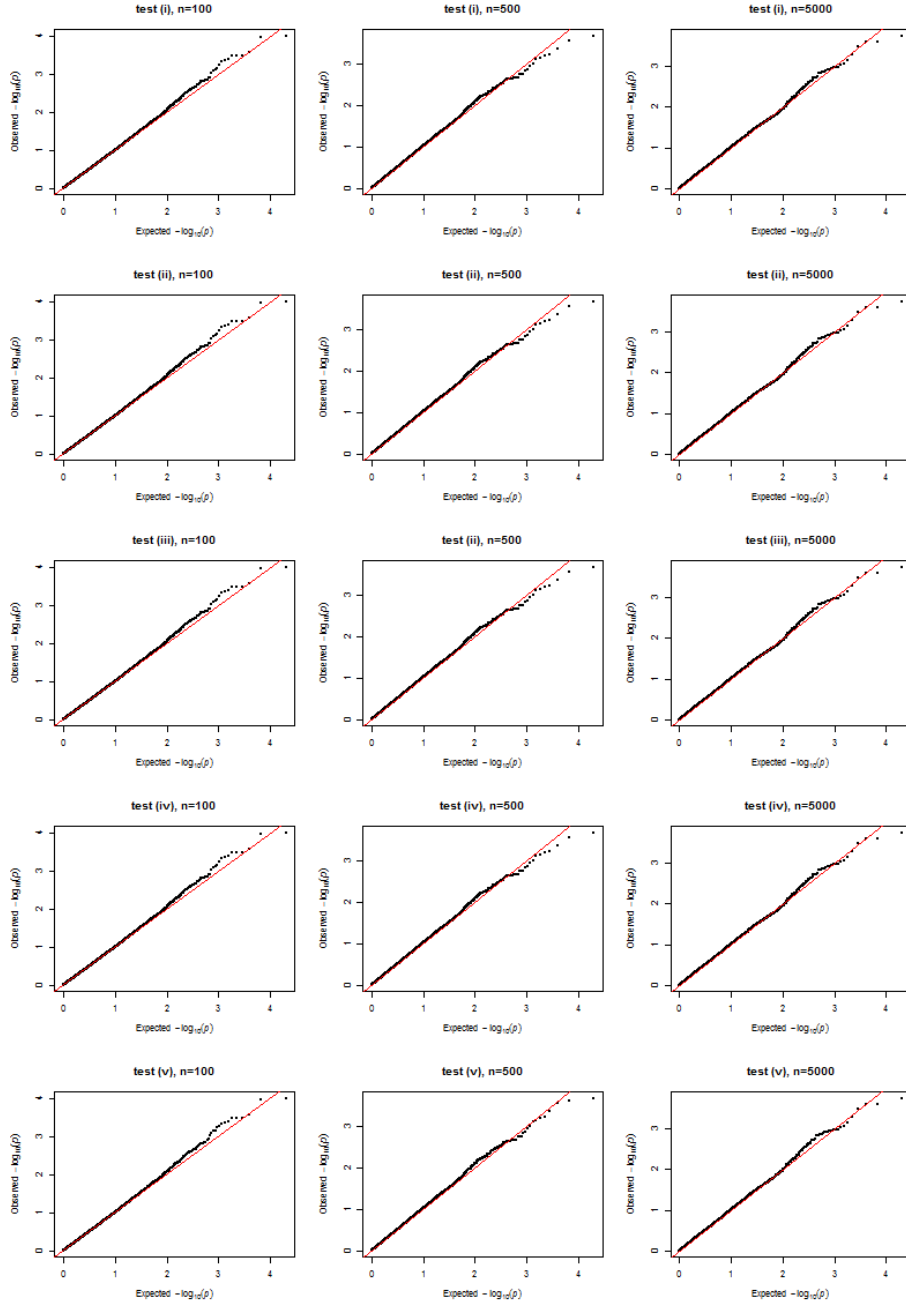


Figure 9. QQ plots of 5th null hypothesis ($\beta_1 = f_1 = \mathbf{0}$) with TPRS

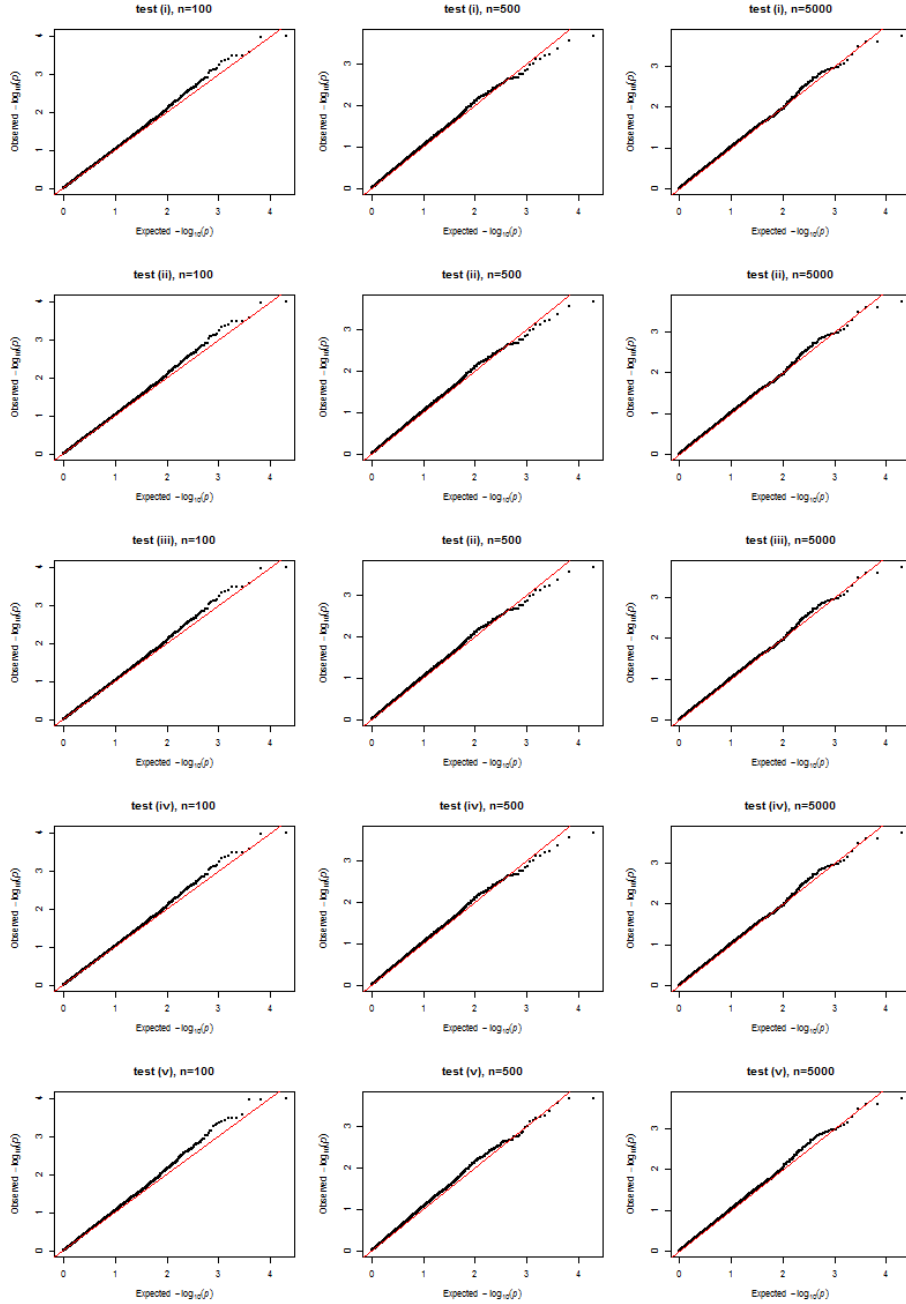


Figure 10. QQ plots of 5th null hypothesis ($\beta_1 = f_1 = \mathbf{0}$) with CRS

Table 4. Simulated p-values which are less than 0.05 and their 95% CIs of 1st null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{0}$)

Obs.	Test	Simulated p-values [95% CI]	
		TPRS	CRS
100	1	0.0661 [0.0611, 0.0710]	0.0654 [0.0606, 0.0703]
	2	0.0662 [0.0614, 0.0711]	0.0656 [0.0609, 0.0706]
	3	0.0659 [0.0611, 0.0709]	0.0644 [0.0595, 0.0693]
	4	0.0660 [0.0612, 0.0709]	0.0650 [0.0602, 0.0699]
	5	0.0770 [0.0717, 0.0822]	0.0817 [0.0762, 0.0871]
500	1	0.0605 [0.0559, 0.0653]	0.0589 [0.0544, 0.0637]
	2	0.0607 [0.0561, 0.0653]	0.0591 [0.0545, 0.0637]
	3	0.0602 [0.0556, 0.0649]	0.0587 [0.0542, 0.0633]
	4	0.0602 [0.0557, 0.0649]	0.0589 [0.0543, 0.0636]
	5	0.0737 [0.0687, 0.0788]	0.0770 [0.0718, 0.0823]
5000	1	0.0596 [0.0551, 0.0642]	0.0568 [0.0522, 0.0614]
	2	0.0597 [0.0551, 0.0644]	0.0572 [0.0527, 0.0618]
	3	0.0594 [0.0550, 0.0640]	0.0564 [0.0518, 0.0610]
	4	0.0594 [0.0549, 0.0641]	0.0566 [0.0521, 0.0613]
	5	0.0719 [0.0669, 0.0769]	0.0760 [0.0708, 0.0811]

Table 5. Simulated p-values which are less than 0.05 and their 95% CIs of 2nd null hypothesis ($\mathbf{f}_1 = \mathbf{f}_2 = \mathbf{f}_3 = \mathbf{0}$)

Obs.	Test	Simulated p-values [95% CI]	
		TPRS	CRS
100	1	0.0722 [0.0673, 0.0773]	0.0781 [0.0728, 0.0834]
	2	0.0723 [0.0673, 0.0774]	0.0784 [0.0732, 0.0837]
	3	0.0720 [0.0669, 0.0770]	0.0778 [0.0725, 0.0832]
	4	0.0720 [0.0671, 0.0771]	0.0781 [0.0728, 0.0834]
	5	0.0884 [0.0828, 0.0940]	0.1018 [0.0959, 0.1078]
500	1	0.0705 [0.0656, 0.0756]	0.0703 [0.0653, 0.0753]
	2	0.0705 [0.0655, 0.0757]	0.0708 [0.0658, 0.0758]
	3	0.0703 [0.0653, 0.0752]	0.0694 [0.0645, 0.0743]
	4	0.0703 [0.0654, 0.0753]	0.0701 [0.0652, 0.0751]
	5	0.0865 [0.0811, 0.0920]	0.0947 [0.0890, 0.1006]
5000	1	0.0638 [0.0591, 0.0688]	0.0652 [0.0604, 0.0701]
	2	0.0639 [0.0592, 0.0687]	0.0656 [0.0607, 0.0705]
	3	0.0634 [0.0586, 0.0683]	0.0651 [0.0602, 0.0699]
	4	0.0635 [0.0588, 0.0684]	0.0655 [0.0608, 0.0704]
	5	0.0800 [0.0748, 0.0852]	0.0878 [0.0823, 0.0935]

Table 6. Simulated p-values which are less than 0.05 and their 95% CIs of 3rd null hypothesis ($\beta_1 = f_2 = 0$)

Obs.	Test	Simulated p-values [95% CI]	
		TPRS	CRS
100	1	0.0584 [0.0538, 0.0631]	0.0584 [0.0539, 0.0631]
	2	0.0586 [0.0540, 0.0632]	0.0586 [0.0540, 0.0632]
	3	0.0583 [0.0537, 0.0629]	0.0580 [0.0534, 0.0626]
	4	0.0583 [0.0538, 0.0628]	0.0582 [0.0536, 0.0629]
	5	0.0639 [0.0591, 0.0688]	0.0676 [0.0627, 0.0726]
500	1	0.0554 [0.0509, 0.0599]	0.0545 [0.0501, 0.0591]
	2	0.0554 [0.0509, 0.0599]	0.0549 [0.0505, 0.0593]
	3	0.0553 [0.0509, 0.0599]	0.0542 [0.0498, 0.0587]
	4	0.0553 [0.0509, 0.0598]	0.0545 [0.0500, 0.0590]
	5	0.0620 [0.0573, 0.0668]	0.0628 [0.0581, 0.0676]
5000	1	0.0572 [0.0527, 0.0617]	0.0549 [0.0505, 0.0594]
	2	0.0573 [0.0529, 0.0619]	0.0554 [0.0511, 0.0599]
	3	0.0569 [0.0524, 0.0616]	0.0547 [0.0503, 0.0592]
	4	0.0572 [0.0527, 0.0618]	0.0550 [0.0505, 0.0594]
	5	0.0625 [0.0580, 0.0672]	0.0637 [0.0591, 0.0685]

Table 7. Simulated p-values which are less than 0.05 and their 95% CIs of 4th null hypothesis ($\beta_1 = f_1 = f_2 = 0$)

Obs.	Test	Simulated p-values [95% CI]	
		TPRS	CRS
100	1	0.0589 [0.0543, 0.0636]	0.0642 [0.0596, 0.0691]
	2	0.0589 [0.0543, 0.0636]	0.0646 [0.0598, 0.0696]
	3	0.0588 [0.0543, 0.0636]	0.0637 [0.0590, 0.0685]
	4	0.0589 [0.0543, 0.0635]	0.0643 [0.0596, 0.0691]
	5	0.0634 [0.0587, 0.0682]	0.0777 [0.0725, 0.0830]
500	1	0.0536 [0.0492, 0.0581]	0.0552 [0.0508, 0.0598]
	2	0.0536 [0.0493, 0.0579]	0.0554 [0.0510, 0.0599]
	3	0.0532 [0.0487, 0.0576]	0.0551 [0.0507, 0.0596]
	4	0.0532 [0.0490, 0.0576]	0.0554 [0.0510, 0.0598]
	5	0.0571 [0.0527, 0.0617]	0.0665 [0.0617, 0.0714]
5000	1	0.0577 [0.0531, 0.0622]	0.0562 [0.0517, 0.0607]
	2	0.0578 [0.0532, 0.0624]	0.0563 [0.0518, 0.0608]
	3	0.0575 [0.0531, 0.0621]	0.0558 [0.0513, 0.0603]
	4	0.0577 [0.0532, 0.0623]	0.0562 [0.0517, 0.0607]
	5	0.0606 [0.0560, 0.0653]	0.0674 [0.0625, 0.0724]

Table 8. Simulated p-values which are less than 0.05 and their 95% CIs of 5th null hypothesis ($\beta_1 = f_1 = 0$)

Obs.	Test	Simulated p-values [95% CI]	
		TPRS	CRS
100	1	0.0515 [0.0471, 0.0559]	0.0537 [0.0493, 0.0582]
	2	0.0515 [0.0472, 0.0558]	0.0537 [0.0494, 0.0582]
	3	0.0515 [0.0472, 0.0559]	0.0537 [0.0493, 0.0581]
	4	0.0515 [0.0472, 0.0558]	0.0537 [0.0492, 0.0581]
	5	0.0522 [0.0479, 0.0566]	0.0577 [0.0532, 0.0623]
500	1	0.0523 [0.0480, 0.0567]	0.0540 [0.0496, 0.0585]
	2	0.0523 [0.0479, 0.0568]	0.0540 [0.0496, 0.0585]
	3	0.0523 [0.0480, 0.0567]	0.0540 [0.0496, 0.0585]
	4	0.0523 [0.0479, 0.0567]	0.0540 [0.0497, 0.0585]
	5	0.0529 [0.0485, 0.0573]	0.0582 [0.0537, 0.0628]
5000	1	0.0490 [0.0449, 0.0533]	0.0502 [0.0460, 0.0545]
	2	0.0490 [0.0448, 0.0532]	0.0502 [0.0460, 0.0545]
	3	0.0489 [0.0447, 0.0533]	0.0502 [0.0460, 0.0546]
	4	0.0490 [0.0447, 0.0533]	0.0502 [0.0460, 0.0546]
	5	0.0498 [0.0455, 0.0541]	0.0521 [0.0478, 0.0566]

2. GEWIS

Table 9 to Table 12 show the effects and their p-values of each SNPs and interactions. For all SNPs, effects in LMM are more significant than in GAMM. However, their total effects are slightly more significant in GAMM than in LMM. The effects of interaction with *pack-year*, moreover, are significant at a critical level of 0.05 in GAMM, but not in LMM.

The main effects of each SNP and their interaction with *smoking-status* variable are similar for both models. Because of smoothness in GAMM, however, effects of variables with smooth function cannot be represented as single value. Therefore, the effects were compared by trends of smoothed variable as Figure 11, which shows the value of estimated smooth function for the variable. As *pack-year* increases, the smoothed interaction of SNP and *pack-year* variable in GAMM tends to monotonically increase, which is similar to the effect of that interaction in LMM.

Table 9. Result of effects and p-values for rs17178251

	LMM	GAMM
Effects of SNP	-0.0249 (2.36×10^{-4})	-0.0179 (0.0328)
Interaction effects of SNP× <i>smoking-status</i>	-0.0285 (0.0459)	-0.0335 (0.0173)
Interaction effects of SNP× <i>pack-year</i>	4.12×10^{-4} (0.1777)	. (0.0341)
Total effects of SNP	. (3.38×10^{-7})	. (1.20×10^{-7})

Table 10. Result of effects and p-values for rs17765644

	LMM	GAMM
Effects of SNP	-0.0248 (2.45×10^{-4})	-0.0180 (0.0324)
Interaction effects of SNP× <i>smoking-status</i>	-0.0289 (0.0424)	-0.0339 (0.0157)
Interaction effects of SNP× <i>pack-year</i>	4.04×10^{-4} (0.1851)	. (0.0373)
Total effects of SNP	. (2.80×10^{-7})	. (9.36×10^{-8})

Table 11. Result of effects and p-values for rs11870732

	LMM	GAMM
Effects of SNP	-0.0251 (2.09×10^{-4})	-0.0183 (0.0292)
Interaction effects of SNP× <i>smoking-status</i>	-0.0276 (0.0536)	-0.0323 (0.0216)
Interaction effects of SNP× <i>pack-year</i>	4.15×10^{-4} (0.1741)	. (0.0361)
Total effects of SNP	. (4.08×10^{-7})	. (1.54×10^{-7})

Table 12. Result of effects and p-values for rs4793541

	LMM	GAMM
Effects of SNP	-0.0242 (3.47×10^{-4})	-0.0172 (0.0401)
Interaction effects of SNP× <i>smoking-status</i>	-0.0288 (0.0437)	-0.0342 (0.0150)
Interaction effects of SNP× <i>pack-year</i>	4.16×10^{-4} (0.1734)	. (0.0343)
Total effects of SNP	. (5.62×10^{-7})	. (1.50×10^{-7})

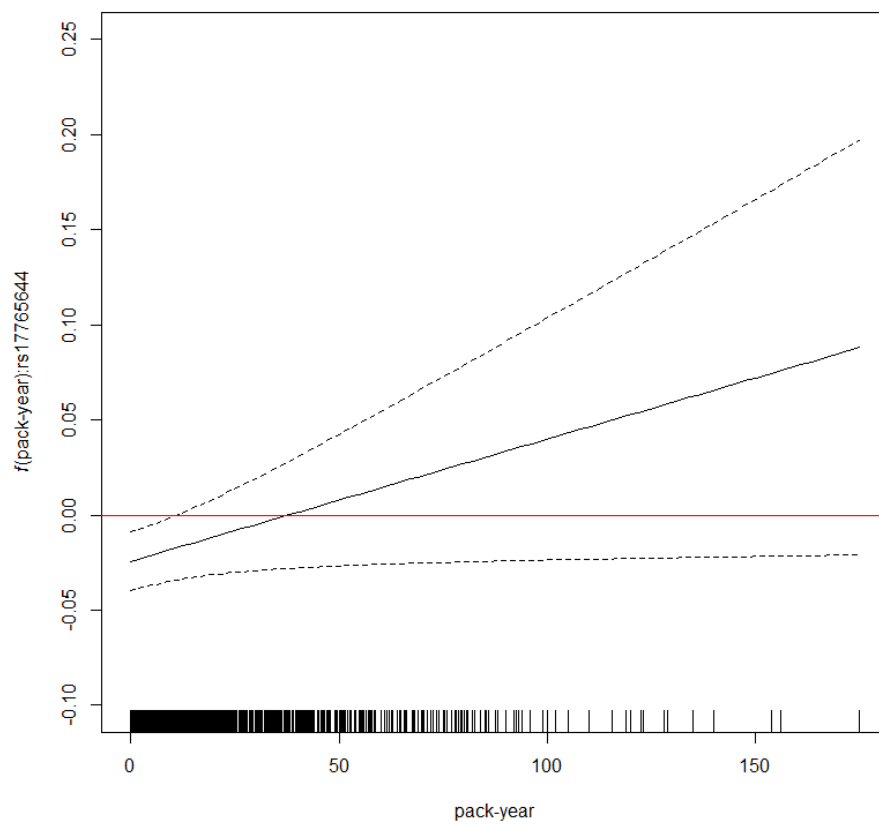


Figure 11. Estimated interaction of smooth function for *pack-year* variable and rs17765644 (Table 10)

IV. Discussion

I extended Wald-type test proposed by S. N. Wood (2013) to various hypotheses, and the proposed methods were evaluated with simulation study. Simulation studies show that the proposed methods preserve the nominal significance level. This improvement will help to make accurate statistical conclusions, especially if p-values are closed to the nominal significance level. Although the effect of the variables cannot be simply expressed as in the linear models, but a best model can be obtained by testing the significance of the smoothed variables or combination of smoothed and linear variables.

Using weighted GAMM in GEWIS for FEV₁, I found that SNPs associated with *SOX9* have significant interaction effect with *pack-year*, even though results from LMM are not. It should be noted that although the weighted model is not entirely consistent with heteroskedastic model, GAM or GAMM are beneficial if there exists non-linear relationship as compared with the linear model. In particular, environmental variables are highly influenced by the nature of the data e.g. race, country, etc., and thus GAM which is more flexible than the linear model would be useful for GEWIS (Cornelis & Hu, 2012). The study in this thesis will also conduct the replication study with additional data in the future.

When various environmental variables are collected, I can fit models that

are more accurate and test the effects of variables by expandability of GAM. However, interpretation is much difficult if the effect of smoothed environmental variables changes complexly, varies with the variables.

The proposed method can be extended to various scenarios with minor modification. Firstly, it is possible to select the best model with GAM from the initial model. The final GAMM used in this thesis is the application of smooth function to environmental variable in the LMM obtained through the model selection process under the linear (mixed) model level. Using the extended Wald-type test for various model structures, the optimal model in the GAM class can be determined.

Secondly, it is necessary to study the relation between the covariance matrix structure and smooth function of GAM. Since the estimation of the covariance matrix is related to the unknown parameter of the systematic component, types of regression spline or the number of bases in smooth function can affect the estimates of covariance matrix. Since there are few related studies, however, fitting GAM with various covariance matrices should be studied. In addition, if the covariance matrix in the linear model is incorporated to the GAM, it should be confirmed whether the statistical validity of the proposed methods is still preserved.

Bibliography

- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421), 9-25.
- Cantoni, E., & Hastie, T. (2002). Degrees-of-freedom tests for smoothing splines. *Biometrika*, 251-263.
- Cornelis, M. C., & Hu, F. B. (2012). Gene-environment interactions in the development of type 2 diabetes: recent progress and continuing challenges. *Annual review of nutrition*, 32, 245-259.
- Crainiceanu, C., Ruppert, D., Claeskens, G., & Wand, M. P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, 91-103.
- Demmel, J. W. (1997). *Applied numerical linear algebra*: SIAM.
- DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, 189-212.
- Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American journal of epidemiology*, 156(3), 193-203.
- Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*, 85-100.
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy.

Statistical science, 54-75.

- Fewster, R. M., Buckland, S. T., Siriwardena, G. M., Baillie, S. R., & Wilson, J. D. (2000). Analysis of population trends for farmland birds using generalized additive models. *Ecology*, 81(7), 1970-1984.
- Gu, C. (2013). *Smoothing spline ANOVA models* (Vol. 297): Springer Science & Business Media.
- Hastie, T., & Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical methods in medical research*, 4(3), 187-196.
- Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (Vol. 43): CRC press.
- Hoffman, K., Webster, T. F., Weinberg, J. M., Aschengrau, A., Janulewicz, P. A., White, R. F., & Vieira, V. M. (2010). Spatial analysis of learning and developmental disorders in upper Cape Cod, Massachusetts using generalized additive models. *International Journal of Health Geographics*, 9(1), 7.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983-997.
- Lancaster, P., & Salkauskas, K. (1986). *Curve and surface fitting*: Academic press.
- Li, L., Zhang, H., Min, D., Zhang, R., Wu, J., Qu, H., & Tang, Y. (2015). Sox9 activation is essential for the recovery of lung function after acute lung injury. *Cellular Physiology and Biochemistry*, 37(3), 1113-1122.

- Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2), 381-400.
- Liu, H., Tang, Y., & Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4), 853-856.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Vol. 37): CRC Press.
- National, C. G. C. U. (2010). Chronic obstructive pulmonary disease: Management of chronic obstructive pulmonary disease in adults in primary and secondary care.
- Parker, R., & Rice, J. (1985). *Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by BW Silverman*. Paper presented at the Royal Statistical Society Series B.
- Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 545-554.
- Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*: Springer Science & Business Media.
- Rockich, B. E., Hrycaj, S. M., Shih, H. P., Nagy, M. S., Ferguson, M. A., Kopp, J. L., . . . Spence, J. R. (2013). Sox9 plays multiple roles in the lung epithelium during branching morphogenesis. *Proceedings of the*

National Academy of Sciences, 110(47), E4456-E4464.

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance components* John Wiley and Sons. *New York, New York, USA*.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*: CRC press.

Suárez-Seoane, S., Osborne, P. E., & Alonso, J. C. (2002). Large-scale habitat selection by agricultural steppe birds in Spain: identifying species–habitat responses using generalized additive models. *Journal of Applied Ecology*, 39(5), 755-771.

Team, R. C. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Vieira, V., Webster, T., Weinberg, J., & Aschengrau, A. (2009). Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. *Environmental Health*, 8(1), 3.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2.

Wahba, G. (1990). *Spline models for observational data*: SIAM.

Webster, T., Vieira, V., Weinberg, J., & Aschengrau, A. (2006). Method for mapping population-based case-control studies: an application using generalized additive models. *International Journal of Health*

Geographics, 5(1), 26.

Wood, S. (2006). *Generalized additive models: an introduction with R*: CRC press.

Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. *R news*, 1(2), 20-25.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95-114.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221-228.

Wu, H., & Lin, J. (2016). A Scaled F Distribution as an Approximation to the Distribution of Test Statistics in Covariance Structure Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3), 409-421.

Young, R. L., Weinberg, J., Vieira, V., Ozonoff, A., & Webster, T. F. (2011). Generalized additive models and inflated type I error rates of smoother significance tests. *Computational Statistics & Data Analysis*, 55(1), 366-374.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5), 2173-2192.

국문초록

일반화가법모형의 평활요소 및 교호작용의 편의 보정 기법에 대한 연구

안 재 훈

서울대학교 보건대학원

보건학과 보건학전공

일반화가법모형 (Generalized additive models, GAM)은 일반화 선형모형 (Generalized linear model, GLM)을 확장한 것으로, 모형의 선형예측식에 평활함수 (smooth function)로 표현된 공변량이 추가되어 모수적 특징과 비모수적인 특징을 모두 가지고 있다. 따라서 반응변수와 공변량의 비선형관계를 보정하기 위한 작업을 생략할 수 있고, 반응변수를 더 정확하게 예측한다. 평활함수에는 다양한 회귀 스플라인 (regression spline)이 사용되며, 적당한 기저 (basis)를 통해 GAM을 penalized GLM으로 접근할 수 있다.

평활함수의 존재로 인해, 일반화가법모형에서의 가설검정으로 얻은 p -값은 특히 기각역에 가까운 경우 부정확한 경향을 나타낸다.

이를 보완하기 위해 S. N. Wood (2013)가 제안한 Wald-type 검정법은 하나의 평활함수에 대한 가설검정에서 그 성능이 입증되었다. 위 논문에서는, 제안된 검정법을 둘 이상의 평활함수 혹은 교호작용에 대한 검정 등 여러 종류의 가설검정에서 적용할 수 있도록 확장했으며, 모의실험을 통해 검정의 정확성이 향상되는 것을 확인했다.

FEV1에 대한 단일염기다형성 (Single Nucleotide polymorphism, SNP)과 환경 요인 변수의 교호작용 분석(GEWS)을 실제 자료로의 적용 대상으로 결정하고, 일반화가법혼합모형 (Generalized additive mixed models, GAMM)과 확장된 검정법을 이용하여 선형혼합모형 (Linear mixed models, LMM)으로 적합한 결과와 비교했다. 분석 대상이 된 4개의 *SOX9* 연관 SNP들은 모두 GAMM에서 평활화된 *pack-year* 변수와 유의한 교호작용을 나타냈지만, LMM에서는 SNP과 *pack-year*의 교호작용이 유의하지 않았다. 여러 자료를 이용한 반복 검증을 통해 교호작용의 효과를 좀 더 명확히 입증할 수 있을 것이다.

일반화가법모형은 환경 요인 변수와 같이 일반적인 선형 관계로 나타내기 어려운 성질을 가지는 변수를 포함한 자료를 분석할 때 장점을 가지고 있다. 평활함수의 효과에 대한 해석과 GAM 수준

에서의 모형 선택 과정 개발 등 몇가지 통계학적 보완이 이루어진
다면 GEWIS를 비롯한 여러 연구에서 정확한 모형 적합과 변수의
유의성 검정에 GAM이 많은 기여를 할 것이라고 기대한다.

주요어 : 일반화가법모형 (GAM), 평활함수, Wald-type 검정법, 전장
유전체-환경 교호작용분석 (GEWIS), *SOX9*

학 번 : 2015-24078